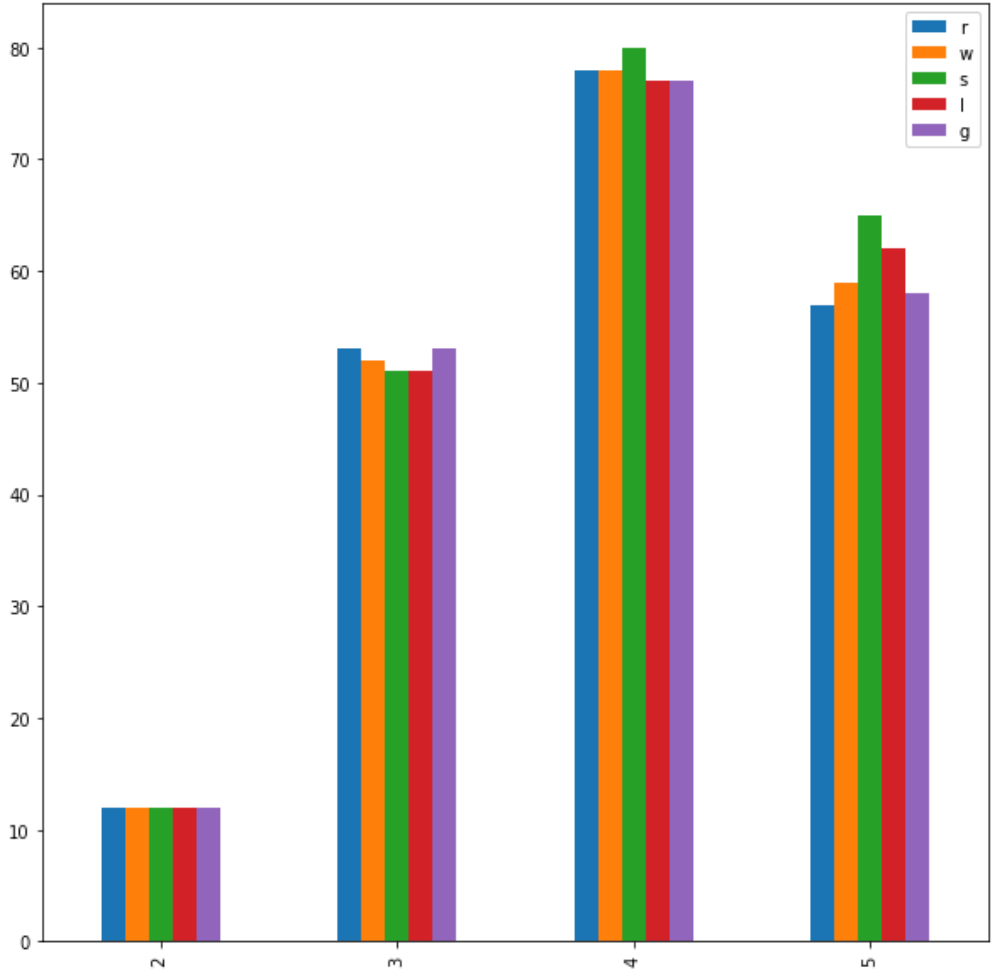
Analysis of bigrams from learners' written work at the Pitt English Language Institute (ELI)

Ben Naismith

Dec 2nd , 2017

Overview of the ELI dataset

- Data collected from 2005 to 2012
- Data collected from speaking, writing, reading, and grammar assignments
- Primarily three levels of proficiency
- Due to length of project, data has taken many forms



Classes Offered

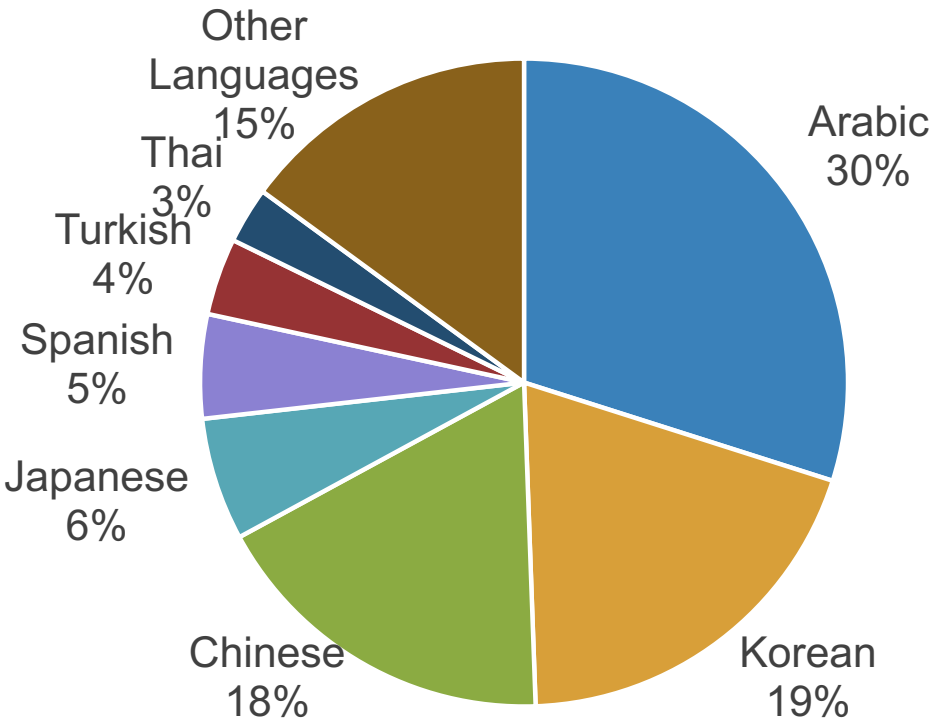
- Reading, Writing, Speaking, Listening, Grammar

- Levels:

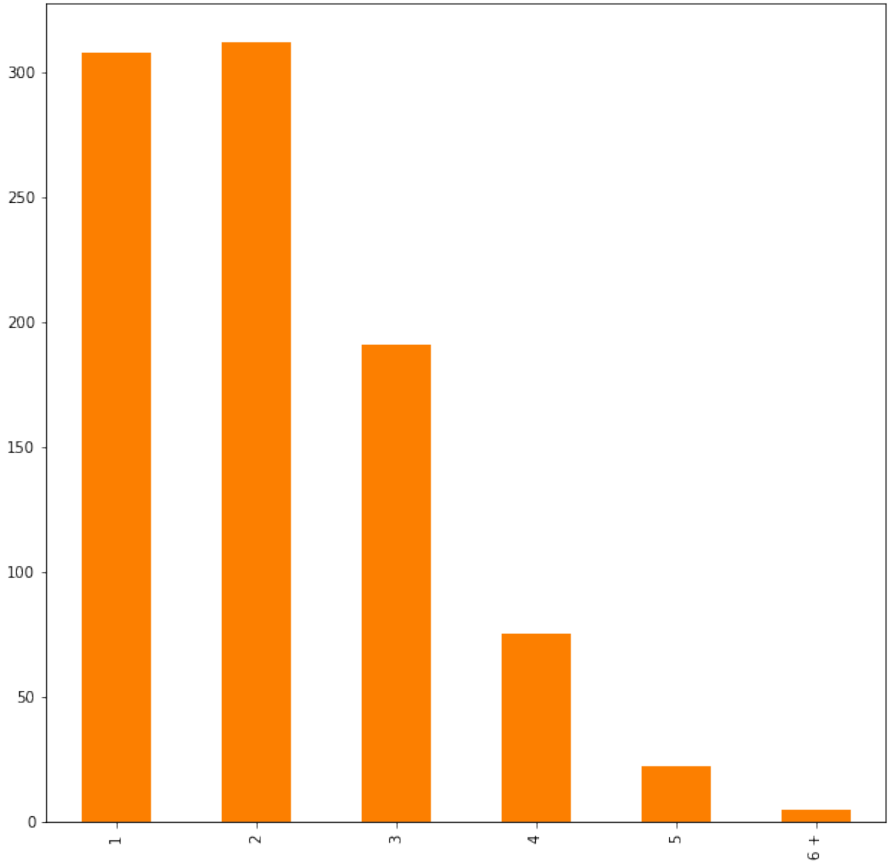
2. Pre-Intermediate
3. Low-Intermediate
4. Intermediate
5. Advanced

Slide: Brianna Hill, 2017

Native Languages



Slide: Brianna Hill, 2017



Semesters Spent at the ELI

- Min: 1 semester
- Max: 8 semesters
- Average: 2 semesters

Slide: Brianna Hill, 2017

Come and visit

https://igx.4sqi.net/img/general/600x600/94064991_tBIvyAxENiwGw0SuOi2GO7DJhYWFYBeKcTDlLZONTqc.jpg

PROJECT OVERVIEW

- familiarize myself with ELI data and help clean/sanitize existing CSV files
- create sub-corpus of written essays for analysis and add statistical information about bigrams
- continue work related to lexical development (cf. Juffs 2015; Juffs 2017), but focusing on the metric of Mutual Information (MI), i.e.
  - *Do bigrams with certain MI scores have noticeable common characteristics?*
  - *Is MI predictive of learner proficiency?*

DATA CLEANING

A big task that the whole ELI Data Mining group was/is working on (some more successfully than others)

Lots of big CSV files!

INDEX.class.csv	separate directories prepared for data stages	2 months ago
INDEX.file_type.csv	separate directories prepared for data stages	2 months ago
INDEX.grammar_point.csv	separate directories prepared for data stages	2 months ago
INDEX.level.csv	separate directories prepared for data stages	2 months ago
INDEX.question_category.csv	separate directories prepared for data stages	2 months ago
INDEX.question_type.csv	separate directories prepared for data stages	2 months ago
answer.csv	separate directories prepared for data stages	2 months ago
course.csv	separate directories prepared for data stages	2 months ago
document.csv	separate directories prepared for data stages	2 months ago
feedback.csv	separate directories prepared for data stages	2 months ago
file_information.csv	separate directories prepared for data stages	2 months ago
grammar_question_pool.csv	separate directories prepared for data stages	2 months ago
question.csv	separate directories prepared for data stages	2 months ago
resource_file.csv	separate directories prepared for data stages	2 months ago
student_information.csv	separate directories prepared for data stages	2 months ago
student_surveys.csv	separate directories prepared for data stages	2 months ago
student_test_scores.csv	separate directories prepared for data stages	2 months ago
user.csv	user.csv file (CONTAINS PRIVATE INFO!!) added.	6 days ago
user_file_internal.csv	separate directories prepared for data stages	2 months ago
user_file_wavtxt.csv	separate directories prepared for data stages	2 months ago

DATA CLEANING (CONT.)

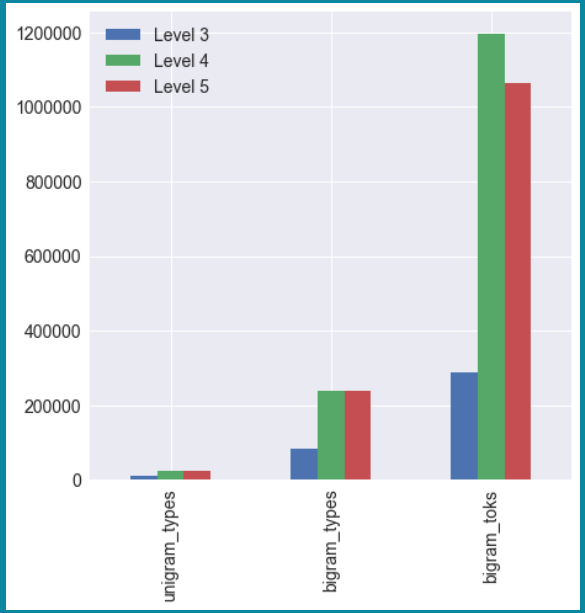
Fun stuff like...

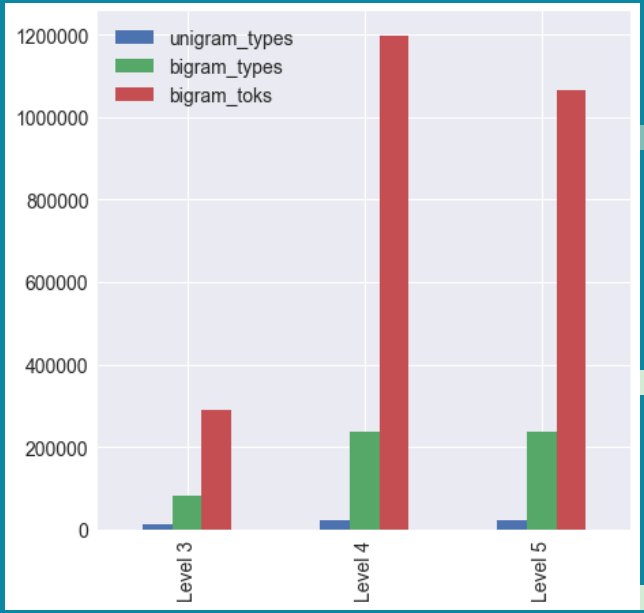
- anonymizing data / removing names
- eliminating fake users
- normalizing codes
- dealing NaN, null, empty strings, etc.
- figuring out how files are linked

DATA ANALYSIS: Big picture

10,956 texts
- 28% level 3
- 39% level 4
- 33% level 5

	unigram_toks	unigram_types	bigram_toks	bigram_types
Level 3	288294	11993	288293	82411
Level 4	1196274	23313	1196273	237014
Level 5	1064445	23753	1064444	237605
Total	2549012	39272	2549011	432880





DATA ANALYSIS: combo_df

answer_id	question_id	user_file_id	anon_id	level_id	course_id	text	toks	bigrams	bigram_len	bigrams_lower	MI_sum	avg_bigram_MI
3	12	7507	dk5	4	115	In my country we usually don't use tea bags. F...	[In, my, country, we, usually, do, n't, use, t...	[(In, my), (my, country), (country, we), (we, ...	67	[(in, my), (my, country), (country, we), (we, ...	181.28	2.71
5	12	7508	ad1	4	115	First, prepare a port, loose tea, and cup.\r\r...	[First, ,, prepare, a, port, ,, loose, tea, ,,...	[(First, ,), (,, prepare), (prepare, a), (a, p...	73	[(first, ,), (,, prepare), (prepare, a), (a, p...	228.84	3.13
7	12	7509	eg5	4	115	First, prepare your cup, loose tea or bag tea,...	[First, ,, prepare, your, cup, ,, loose, tea, ...	[(First, ,), (,, prepare), (prepare, your), (y...	49	[(first, ,), (,, prepare), (prepare, your), (y...	120.11	2.45
8	13	7509	eg5	4	115	I organized the instructions by time, beacause...	[I, organized, the, instructions, by, time, ,,...	[(I, organized), (organized, the), (the, instr...	38	[(i, organized), (organized, the), (the, instr...	102.94	2.71
11	12	7511	fv6	4	115	To make tea, nothing is easier, even if someti...	[To, make, tea, ,, nothing, is, easier, ,, eve...	[(To, make), (make, tea), (tea, ,), (,, nothin...	98	[(to, make), (make, tea), (tea, ,), (,, nothin...	269.31	2.75

Mutual Information (MI)

- Strength of the association between words, i.e. the two-way likelihood of them co-occurring (Simpson-Vlach & Ellis, 2010)

- Statistical measure which corresponds most closely to native speaker judgements of the salience of formulaic sequences (Paquot & Granger, 2012)

- *Let's get interactive*
- *What about a general corpus?*

DATA ANALYSIS: bigram_df

	bigram	tokens	MI	per_million	lv3_norm_toks	lv4_norm_toks	lv5_norm_toks	level_3	level_4	level_5	lv3_per_M	lv4_per_M	lv5_per_M
1	[in, my]	2629	3.13	1031.38	621	1184	822	23.65%	45.06%	31.29%	243.62	464.49	322.48
2	[my, country]	875	5.50	343.27	229	341	303	26.27%	39.08%	34.65%	89.84	133.78	118.87
3	[country, we]	17	0.36	6.67	1	11	3	11.69%	65.10%	23.21%	0.39	4.32	1.18
4	[we, usually]	80	3.41	31.38	12	53	13	15.81%	67.01%	17.18%	4.71	20.79	5.10
5	[usually, do]	53	3.07	20.79	5	28	19	9.58%	53.00%	37.41%	1.96	10.98	7.45

DATA ANALYSIS – 'Top 20 Tables'

- *Bigram tokens*
- *High MI*
- *'Medium' MI*
- *Bigram tokens per level*
- *Bigrams indicative of level*

MACHINE LEARNING!

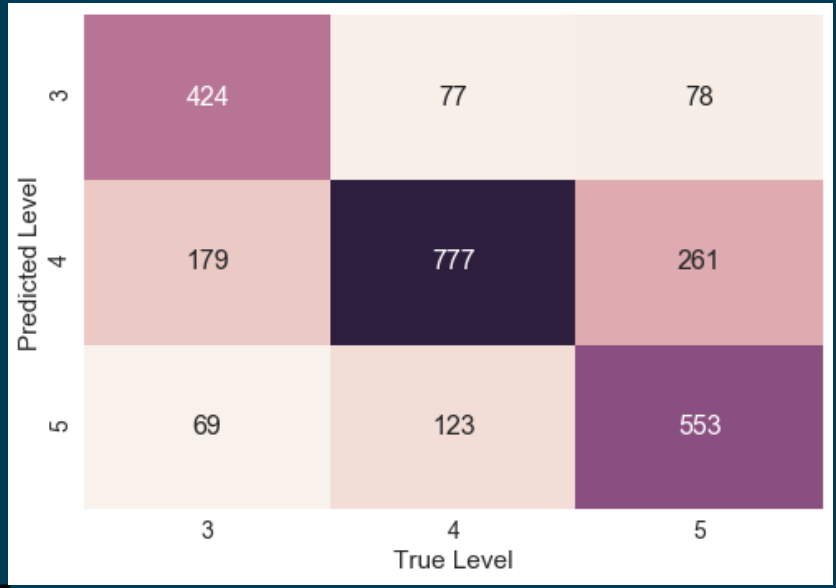
(my arch-nemesis)

- Predict level based on words used
- Predict level based on average MI of every bigram in a text

see Visualizations.ipynb

MACHINE LEARNING (CONT.)

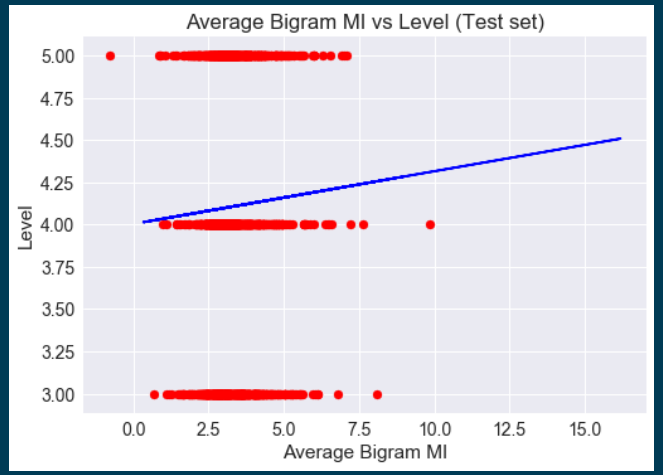
Predict level based on words used:

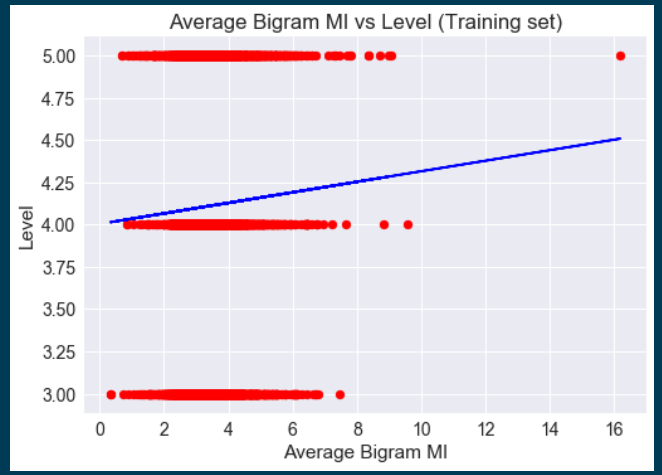


Pretty good accuracy at 69% although this may have to do with the lexical sets of the prompts.

MACHINE LEARNING (CONT.)

Predict level based on average MI of every bigram in a text





No significant difference between average overall MI and level

IMPLICATIONS

- Overall average MI of a text is not predictive – perhaps due to even increase of grammatical words and meaningful collocations as level increases (hypothesis needs to be tested)

- Bigrams with very low MI (uncommon/grammatical) and very high MI (compound nouns, proper names) are not as useful for learners/teachers – what is the 'sweet spot'?

WHAT I'VE LEARNT

- Challenges of working with real, i.e. messy data
- Benefits of working with others in corpus research
- Possibilities of corpus research without resorting to GUIs
- Many new ideas for future research

POSSIBLE FUTURE RESEARCH

- Finding which range (if any) of MI in lexical items predicts usefulness (judged by humans) and level (machine learning)
- Comparative studies with general, non-learner corpora
- Comparison of these findings with spoken texts
- Investigating any differences between L1s

REFERENCES

Juffs, A. (Forthcoming). Lexical Development In The Writing Of Intensive English Program Students. In R.M. DeKeyser & G. Preito Botana (Eds.), *Reconciling methodological demands with pedagogic applicability*. Amsterdam: John Benjamins.

Juffs, A. 2017. The longitudinal development of lexical bundles in the written output of Arabic-speaking ESL learners. Unpublished manuscript.

Paquot, M., & S. Granger. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*:130-149. doi:10.1017/S0267190512000098

Simpson-Vlach, R. & N.C. Ellis. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics, 31*(4): 487–512. doi:10.1093/applin/amp058

THANK YOU!

Any questions?

You can contact me at:
@bennaismithelt
bnaismith@pitt.edu